SUPeRB: Dealing with bibliographic references in the Portuguese-speaking world

Luís Miguel Cabral, Diana Santos, Luís Costa
{Luis.M.Cabral, Diana.Santos,Luis.Costa}@sintef.no

Linguateca

Presentation Outline

- Background
- Motivation
- SUPERB
  - Features
  - Use Cases
  - Architecture
  - Evaluation
- Current status
- Future Work
- Conclusions

Background

- Linguateca is a distributed resource center for Portuguese
  - http://www.linguateca.pt/
- The Linguateca's Catalogue of publications aims at providing bibliographic information about Portuguese NLP
  - Gathering about 1547 references
  - 1116 links for the electronic version of the document (paper or presentation)
  - 988 total publications with a digital document link

Motivation

- Ease access to documentation written within a specific domain
- Improve an existing resource for the Portuguese speaking community, the Linguateca's publication catalogue
  - Improve its usability
  - Provide better means of gathering information from the web sources
- Improve processing of references written by Portuguese native speakers

SUPeRB

Sistema Uniformizado de Pesquisa de Referências Bibliográficas

Is an semi-automatic system which helps in the search and the managing of bibliographic information
- Extraction of new bibliographic information
- Management of the bibliographic information within a catalogue
  - Add and edit
  - Tag
  - Validate
  - Generate summary pages

Features

- Capable of:
  - Search and get bibliographic sources existing on the web
  - Process several digital document formats, converting them to a common format
  - Identify relevant information
  - Extract bibliographic details
  - Support edition and validation of bibliographic information
  - Integrate the information in a bibliographic catalogue
SUPeRB: Some terminology

- Bibliographic references and elements
- Bibliographic formats (BibTeX, RIS, EndNote)

Using SUPeRB

- Two kinds of users
  - Repository user, allowed to search, suggest and tag bibliographic references
  - Repository manager, responsible for validating users actions and generate final data

- Input
  - Expression composed by a set of keywords
  - Partial reference
  - Link to a web document

- Output
  - Relevant bibliographic content, properly formatted for use or storage

Architecture

WebSearch
Searching the Web

- Module responsible for processing expressions given by users, rewriting them and generating multiple expressions
  - Joins specific domain words into the expressions,
  - Queries search APIs (Google API, Yahoo API)
  - Returns URLs to relevant digital documents on the Web

DocumentHandler
Convert documents into text

- Obtain the content from different file formats
  - Gets remote documents if necessary
  - Converts documents formats into text
    - Using publicly available programs
    - Can be easily configured to use other programs

ReferenceExtractor
Finding references from unstructured text

- Capable of identifying and delimiting bibliographic references
  - Relies first on classification methods to match the document structure to expected genres (article, dissertation, homepage, ...)
  - Handles each structure in a different way
    - Article's header
    - Article's reference section
    - List of references
ReferenceParser
 Parsing bibliographic references
- Takes one bibliographic reference and obtains its bibliographic elements (author, date, etc.), using a combination of methods
  - Paratools (Jewell)
  - Heuristics
  - Ontology based validation (REB) before user validation
- REB (Repositório de Elementos Bibliográficos) is both:
  - An ontology of authors, conferences and places allowing relations between elements (dealing for instance with abbreviations)
  - A set of similarity detection methods (both used for updating the ontology and validating the reference)

ReferenceMerger
 Avoiding duplicates
- Identifies duplicates or partial references (different parts of the same reference)
- Merges a set of partial references, combining the distinct elements in each
- Checks whether the reference is already in the catalogue and if it needs updating

Other Modules
- ReferenceTagger
  - Allows users to provide additional information by assigning tags
- ReferenceConverter
  - Provides conversion between different bibliographic formats (BibTeX, EndNote, etc.)

Evaluation of ReferenceParser
- Followed a methodology inspired by HAREM
  - 33 bibliographic references in several languages in the NLP area
  - 239 bibliographic elements
  - Each element can be:
    - Correctly delimited and classified
    - Correctly delimited but misclassified
    - Correctly classified but incorrectly delimited (partially correct)
    - Missing

These results helped detect several bugs which have now been fixed

Current status
- SUPeRB manages more than 2000 publications ➔ 120 virtual places of publication (conferences, books and journals)
- SUPeRB produces aggregated reports in several languages and generates the corresponding charts
  - places, dates
  - conventions
  - sorting
Future Work

- Integrate faceted search with SUPeRB
- Extract keywords from abstract or the complete text
- Accept user defined ontologies
- Scheduling for publications related tasks
  - checking missing fields
  - making available electronic versions

Availability

http://www.linguateca.pt/SUPeRB/

- Manages Linguateca's catalogue
- Source code available separately
- Documentation in Portuguese

Concluding remarks

- Motivated by a practical problem
- System used in practice
- An example of information extraction of a specific kind of information (references)
- Potential usefulness in virtually any scientific area with Portuguese speaking authors
- Publicly available

Acknowledgements

This work was done in the scope of the Linguateca, contract nº39/1.3/C/NAC, project jointly funded by the Portuguese Government and the European Union.